決定リストによる動詞語義曖昧性の解消

1Q－6

今西 奈美　成嶋 弘　峯崎 俊哉
東海大学

1 はじめに

自然言語処理において、単語の語義曖昧性解消は重要な問題の一つである。そこで、本稿では、動詞に焦点を絞り、決定リストを適用した語義の曖昧性解消方法について報告する。この方法は、音声合成における同形異音語の読み分けを行う為に提案された方法である。Yarowsky [1] は、同形異音語の読みを判断できる「証拠」をコーパスから自動的に獲得し、証拠と証拠に対する読みと評価値をリストにし、それらを決定リストと呼んでいる。また、Li ら [2] は、決定リストの信頼性を考慮し、信頼できない証拠は決定リストから削除することにより、統計的に信頼できる決定リストを用いて、高い正解率で読みの判断を行うことに成功した。本研究は、同形異音語の読みの判断を行う方法を、動詞の語義の判断に応用したものである。

2 決定リスト

2.1 語義判定のための証拠

ⓐ 直前、直後の品詞
ⓑ 直前、直後の単語
ⓒ 直前5単語、直後5単語内に共起する助詞
ⓓ 直前数単語、直後数単語内に共起する単語

Verb Sense Disambiguation Using Decision List
Nami Imanishi, Hiroshi Narushima, Toshiya Minezaki
Tokai University
1117 Kitakaname, Hiratsuka, Kanagawa 259-12, Japan

2.2 決定リストの学習

語義を判断するための証拠を学習データ（EDR コーパスを使用する）から以下の方法で獲得し、決定リストを作成する。

① 各証拠に対して、その証拠の下での語義の条件付き確率 $P(D|E)$ を推定する。この時、確率変数 D は語義を表す。確率変数 E は証拠が語義と共起するかしないかを表す。(共起するときは1，共起しないときは0をとる)
② 各証拠に対して、その証拠の下での確率の値が上位二つの語義 d_i , d_j ($\hat{P}(D=d_i|E=e) \geq \hat{P}(D=d_j|E=e)$) をみつける。そして、次の式によって各証拠に与えられる確率比を，各証拠の「尤度比」と呼ぶ。

$$L(D=d_i|E=1) = \log_2 \frac{\hat{P}(D=d_i|E=1)}{\hat{P}(D=d_j|E=1)}$$

③ すべての証拠を尤度比の降順にソートする。
④ 絶対出現確率の値が上位二つの語義 d_i , d_j ($\hat{P}(D=d_i) \geq \hat{P}(D=d_j)$) の確率比を次の式によって与える。

$$L(D=d_i) = \log_2 \frac{\hat{P}(D=d_i)}{\hat{P}(D=d_j)}$$

⑤ 尤度比が④で求めた値より小さい証拠を削除する。また，必ず語義が選択されるように④で求めた語義と確率比を決定リストに付け加える。
⑥ 以上の方法で得られた決定リストの信頼性を考慮するために、各証拠に対して、その証拠と動詞間の相互情報量を次の式によって計算する。

$$I(D,E) = \sum_{d\in D, e\in E} P(d,e)\log_2 \frac{P(d,e)}{P(d)P(e)}$$

⑦ 閾値を次の式によって定める。

$$\theta(D,E) = \beta \cdot \frac{(K_E - 1)(K_D - 1) \cdot \log_2 N}{2 \cdot N}$$

N : データ数
K_E: 変数 E の取りうる値 (すなわち 2)
K_D : 変数 D の取りうる値
β : $0 \leq \beta \leq 1$

相互情報量 $I(D,E)$ が閾値 $\theta(D,E)$ 以下であれば、その証拠を決定リストから削除する。

3 実験

本研究では、動詞をランダムに選び、それらの動詞を含む文を集め、それぞれデータセットを作成する。実験は、5 fold Cross Validation の形で行う。すなわち、抽出したデータセットを均等に 5 分割し、80% のデータセットを訓練データとして決定リストを作成し、残りのデータセットをテストデータとし正解率を計算する。さらに、テストデータの選び方を変えて，これを 5 回繰り返して、正解率の平均を求める。

比較対照の為、決定リストを使用せず共起確率のみで語義を判断する実験 (実験 1)、データの信頼性を考慮しない決定リストによる語義判断の実験 (実験 2)、本研究で提案する実験 (実験 3) を行った。ただし、2.1 の⑷の証拠については，(1) 直前直後 5 単語内に共起する名詞、(2) 直前直後 5 単語内に共起する単語、これら二つの場合に対して実験を行った。以下に実験結果を示す。(実験 3 においては，$\beta = 0.2$ とした。)

証拠 ⑷	実験 1(%)	実験 2(%)	実験 3(%)
(1)	82.42	82.83	83.50
(2)	76.93	76.97	76.91
平均	79.68	79.90	80.20

また、実験 2 と実験 3 の決定リストの長さの平均を比較した結果、次のとおりである。

実験	平均
実験 2	182.43
実験 3	100.81
縮小率	44.74(%)

以上の実験結果より、証拠に直前直後 5 単語内に共起する名詞を使用し、本研究で提案した方法により語義を判断した方が、よい結果を得られることがわかった。また、信頼性を考慮することにより、少ない証拠でも、高い正解率で語義を判断できることもわかった。

4 おわりに

本稿は、データの信頼性を考慮した決定リストによる動詞語義曖昧性の解消法を提案した。今後、さらに精度をあげるために、どのような証拠を使用すればよいのか等の工夫を行うことが必要である。

謝辞

本研究の方法を示唆していただいた，NEC C&C 研究所 Hang Li 氏に感謝いたします。

参考文献

[1] David Yarowsky : Decision lists for lexical ambiguity resolution : Application to accent restoration in Spanish and French. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. pp.189-196 (1995)

[2] Hang Li Junichi Takeuchi : Using Evidence that is both Strong and Reliable in Japanese Homograph Disambiguation
情報処理学会自然言語研究報告 119-9 (1997)